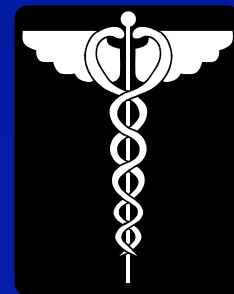


# Survival Analysis

Joe Schwartz

Department of Psychiatry and Behavioral Sciences  
Applied Behavioral Medicine Research Institute  
SUNY - Stony Brook



# Concepts to be covered

1. Survival curve/function
2. Hazard rate
3. Censored data
4. Kaplan-Meier estimate of survival distrib
5. Hazard ratio
6. Proportional hazards assumption
7. Comparing survival curves
8. Proportional hazards regression analysis
9. Time-varying covariates

# When is Survival Analysis used?

- When the outcome of interest is the occurrence of an event, and you have information on:
  - Whether the event occurred, and
  - When the event occurred (time or date)
- Sometimes called “Time-to-event” data

# Examples of outcomes

- Death
- Heart attack and/or CV surgery
- Smoking first cigarette during a quit attempt
- Birth of a child (maternal age or pregnancy duration)
- First use of an illicit drug
- Hospitalization for a psychiatric disorder

# Inappropriate outcomes

- Birth weight, blood pressure - not events
- Smoking status - not an event
- # cigarettes smoked/day - not an event
- Birth by C-section - event, but not timed
- Whether patient/subject takes his/her prescribed medications, brushes teeth in the morning, or has ever tried to quit smoking - all events, but outcomes are not linked to the passage of time

# Questionable outcomes

- Onset of alcohol dependence or various chronic diseases (Type II diabetes, fibromyalgia, rheumatoid arthritis, atherosclerosis) - often difficult to date the onset

# Survival function

## heuristic conceptualization

Suppose you do research on fruit flies and you have a jar filled with 1000 male adult fruit flies.

Suppose, also, that every day 3% of your fruit flies die; none are born (all males!).

What happens to your population of fruit flies over the course of a month?

# Survival function

## heuristic conceptualization

At the end of day one, 30 fruit flies have died, leaving 970 alive.

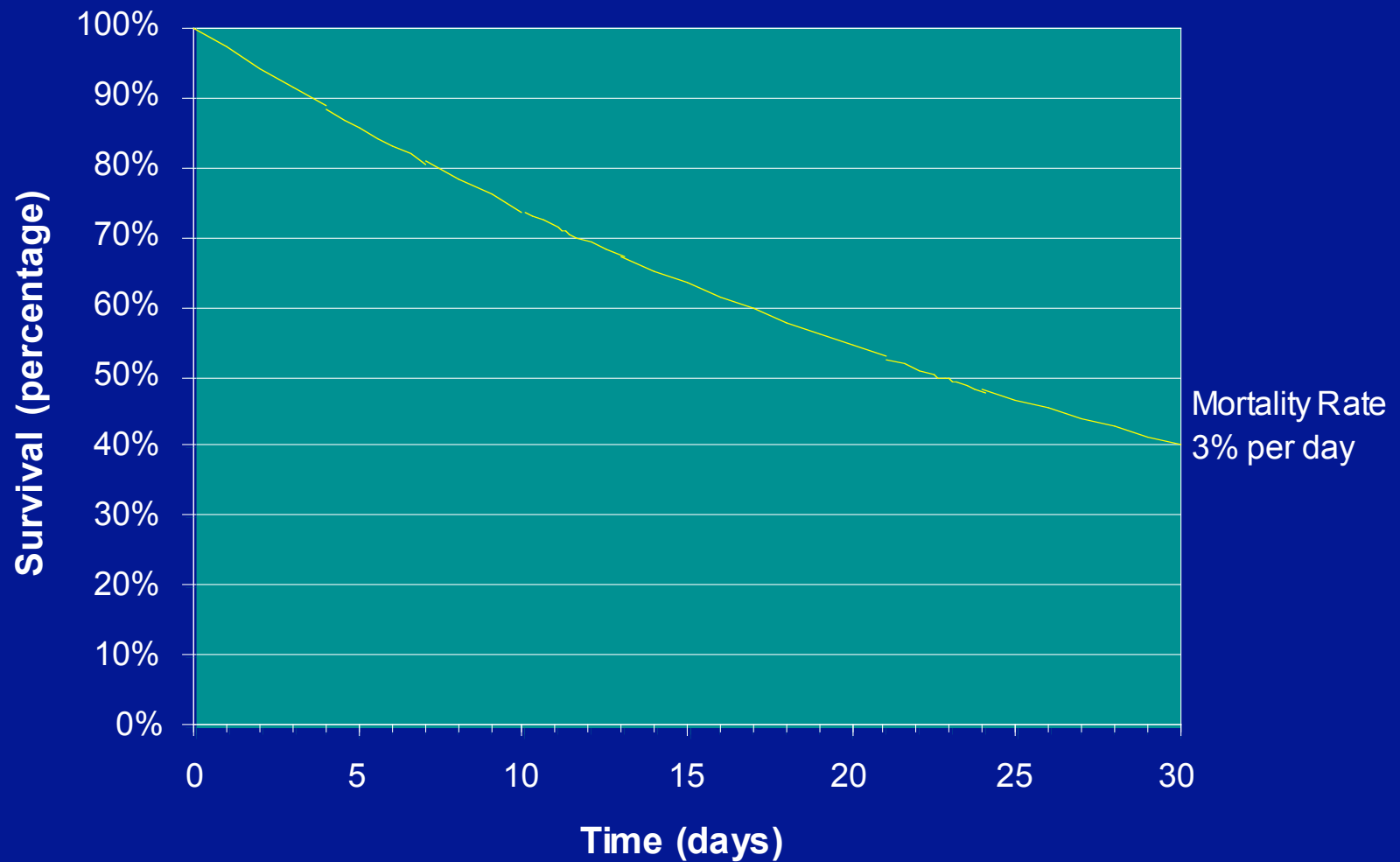
During day 2, 3% of 970 (=29) fruit flies die, leaving 941 at the end of the day.

During day 3, 3% of 941 (=28) fruit flies die, leaving 913.

.....

During day 30, 3% of 413 (=12) fruit flies die, leaving 401.

## Example of Survival Curve when Hazard Rate is Constant



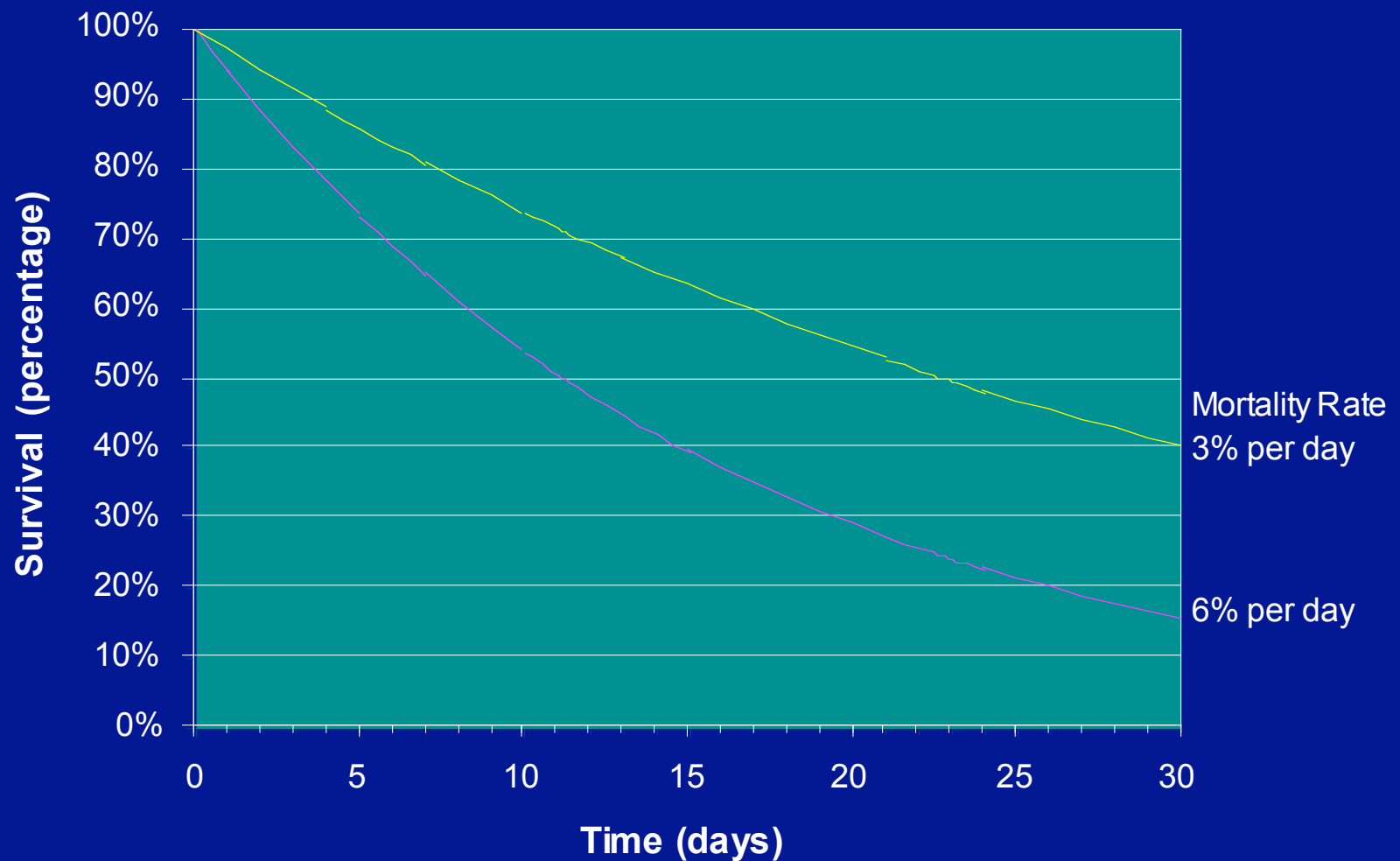
# Survival function

Survival is a function of time - the more time that elapses the smaller is the probability that the event has not occurred

The survival function gives the probability, for each value of  $t$  (time), of the outcome event not having occurred

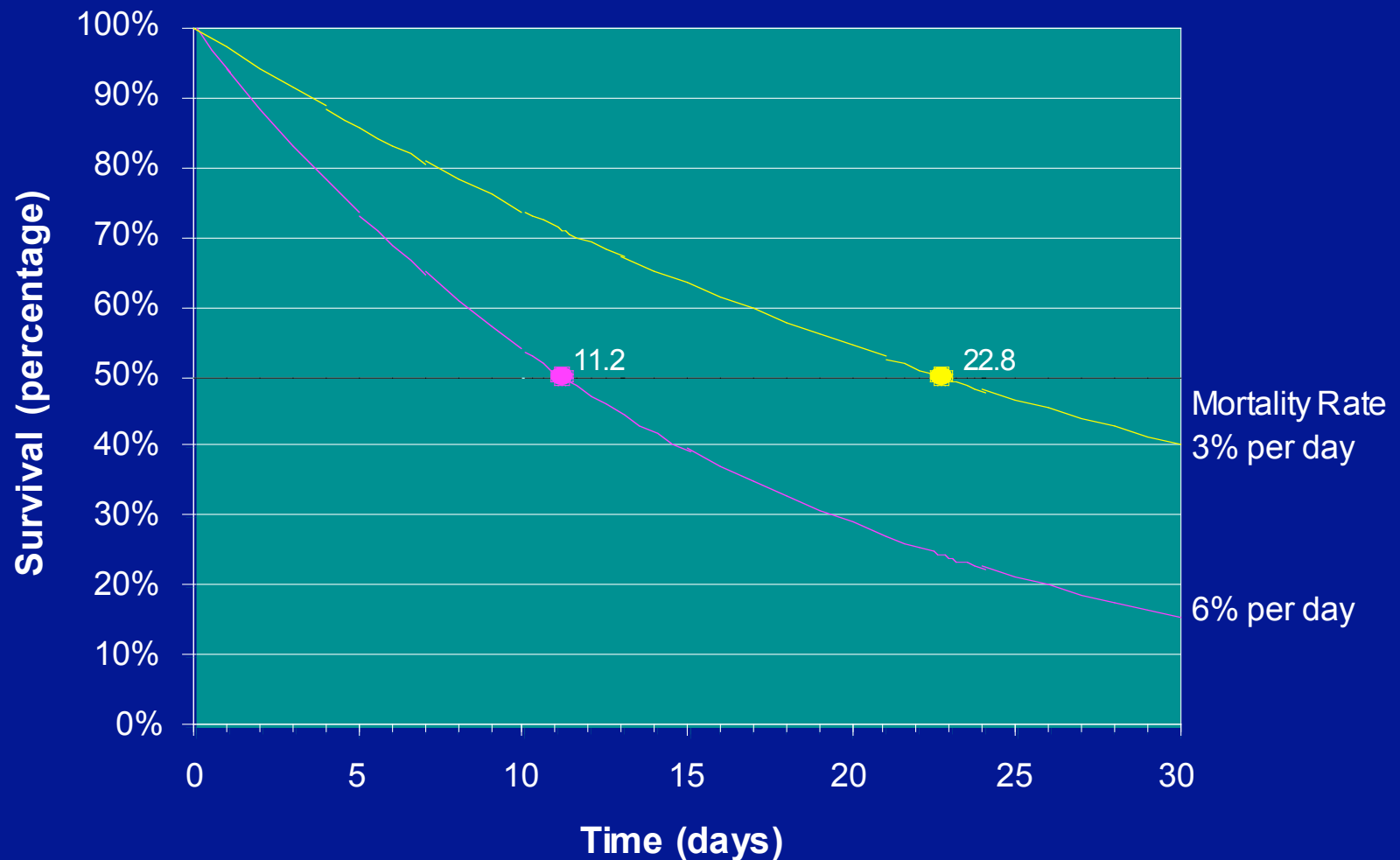
$$S(t) = \text{Pr}(\text{event } \underline{\text{not}} \text{ occurring prior to time } t)$$

## Examples of Survival Curves when Hazard Rate is Constant

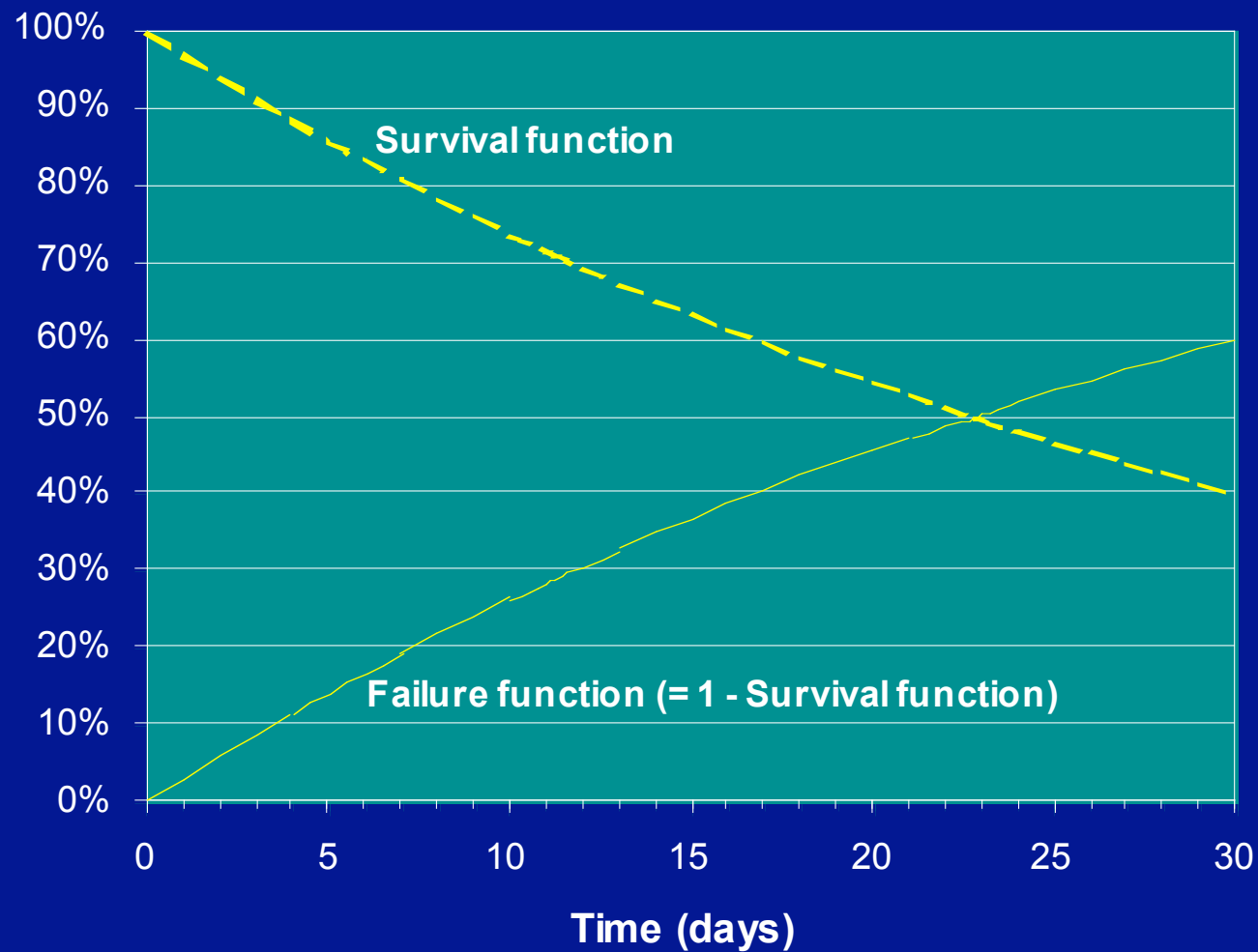


## Examples of Survival Curves when Hazard Rate is Constant

### Median Survival Time



## Example of Survival Curve and Corresponding Failure Curve



# Where are we?

- You've seen a couple of survival functions (for hypothetical data with a constant event rate)
  - survival curves are monotonically decreasing (they can never curve upwards)
  - median survival time: equals the time when 50% of the population have had the event
- Failure function is “the complement” of the survival function:  $F(t) = 1 - S(t)$

# Hazard Rate

The hazard rate is the instantaneous rate at which events occur (*pretty abstract*)

# Hazard Rate

Analogy: Think about a savings account in which you deposit \$1000 and leave it for 1 year. The bank says its interest rate (APR) is 4%/yr. If they “compound interest daily (or continuously)” then after one year, you will actually earn \$40.81 interest (APY = 4.081%/yr), due to earning interest on interest.

If APR=20%, then APY=22.14%.

The Hazard Rate is analogous to the APR.

# Hazard Rate

In the earlier survival function for a constant mortality rate of 3%/day, the hazard rate was 3.05%/day. For the survival function with a constant mortality rate of 6%/day, the hazard rate was 6.19%/day. (the hazard rate is always a little more than the actual event rate)

# Relationship of Survival Function to Hazard Rate

When the hazard rate ( $\lambda$ ) is constant, then

$$S(t) = e^{-\lambda t}$$

Note that  $S(t)$  equals 1.00 (i.e., 100% survival) when  $t=0$

$S(t)$  is an “exponential function” implying that the distribution of survival times has an “exponential distribution”

Median survival time =  $-\ln(0.5) / \lambda = 0.693 / \lambda$ , and

$$\lambda = -\ln(0.5) / \text{median survival time}$$

# Hazard functions

In the above examples, we have assumed that the hazard rate is a constant, perhaps a different constant for different groups.

However, this is a very restrictive assumption.

Other models assume that the hazard rate changes, usually as a function of time  $[\lambda(t)]$ ; perhaps increasing over time or decreasing over time. Two extensions of the exponential model are the Weibull model and the Gompertz model.

# Examples of events where hazard rate changes over time

All cause mortality: is elevated during first year of life, drops and remains quite low through about age 30, and then increases steadily over the rest of the lifespan.

Initiation of smoking: very low during pre-adolescence, increases through teen years, and decreases markedly after about age 20; if one hasn't started smoking by age 30, the chances of becoming a smoker are very low.

# Why should you care about hazard rates?

Because they are considered the “fundamental parameter” or “driving mechanism” for the survival function: the survival distribution is what we can observe, but it is the hazard function that we want to model.

Again, it is analogous to an interest rate: you can watch your savings account grow, but if you want to understand “the process” by which it grows, you have to know the interest rate. (Another parallel: interest rates also change over time.)

# A real example

- A smoking cessation study designed to:
  - test whether individuals randomly assigned to a structured social support program are more successful at quitting than those who receive general advice about quitting but are not provided the structured support program.
  - Duration of study: 3 weeks (1 week prior to start of quit attempt and 2 weeks after)
  - Primary outcome: time from the start of the quit attempt until participant first smokes a cigarette

# Alternative analyses

- Could use a chi-square test (or logistic regression) to test whether those in the treatment group (T) are more likely to go 14 days without smoking, a binary outcome, than those in the control group (C).
- Could test whether the average (or median) time from quitting until the first cigarette is longer in group T than group C.
- Issue: censored observations

# Censored observations

- For both types of analysis, how do you treat someone who drops out of the study after 8 days, without having yet smoked?
- When time until first cigarette is the primary outcome, what value do you use for those who have still not smoked at the end of the 14-day observation period? 14? 15? Missing data?

# Survival function - a real example

- Distribution-free estimation of survival curve
  - Kaplan-Meier (aka “product limit”) method
  - handles censored data

**Raw data** for the control group in a smoking cessation study.

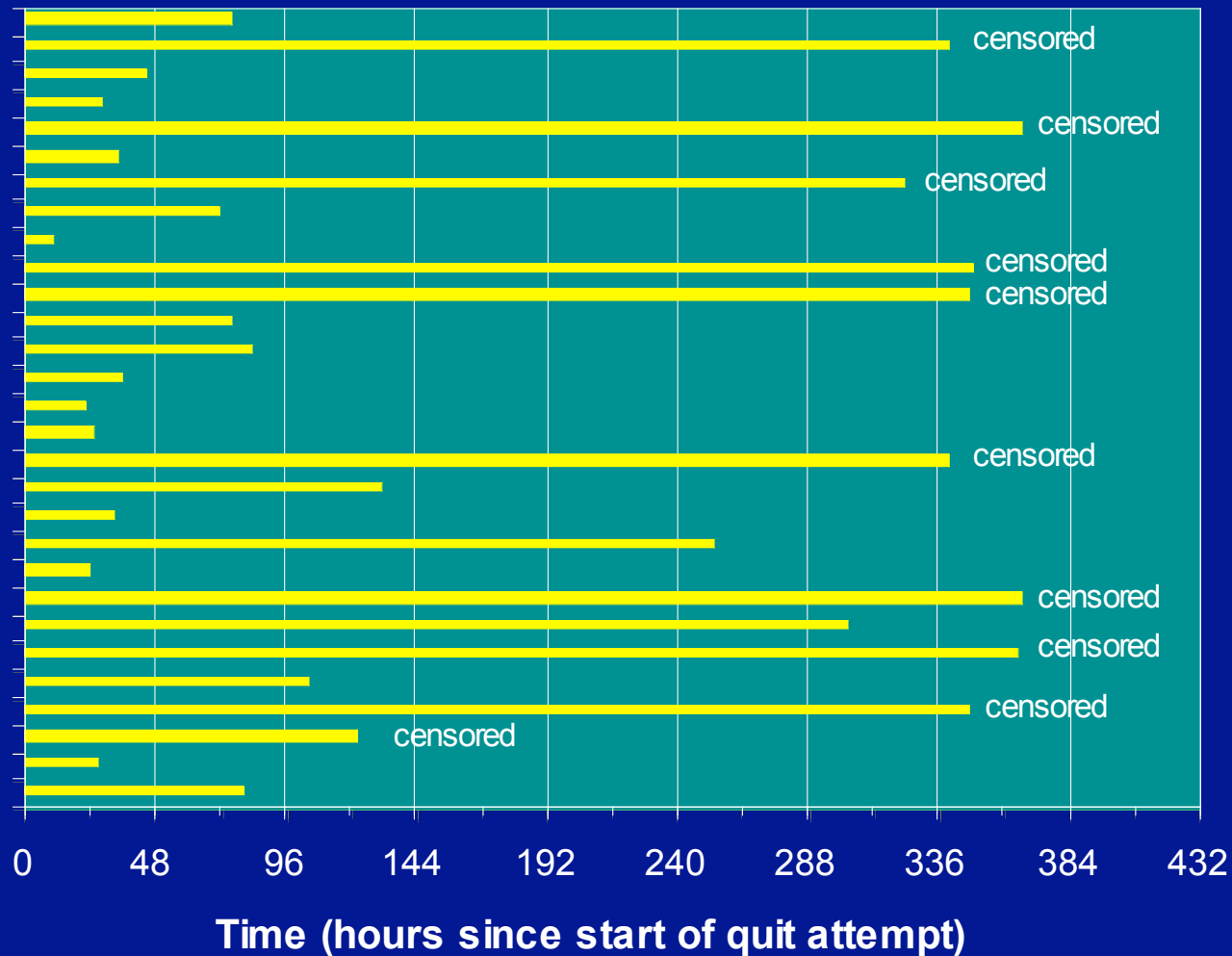
N = 29

Hours: number of hours from beginning of quit attempt until first cigarette (lapse) or observation “censored”

Censored: last observation of subject, at which time the target event has not occurred (yet?)

<u>ID</u>	<u>Hours</u>	<u>Status</u>
307	80.0	lapse
309	26.1	lapse
318	122.4	censored
325	346.8	censored
326	103.3	lapse
327	365.3	censored
328	302.4	lapse
329	366.1	censored
331	23.8	lapse
333	252.7	lapse
337	32.4	lapse
338	130.6	lapse
341	340.0	censored
352	24.3	lapse
354	21.6	lapse
359	34.8	lapse
364	82.4	lapse
365	75.5	lapse
366	346.9	censored
370	347.6	censored
373	10.5	lapse
376	70.9	lapse
380	322.4	censored
384	33.3	lapse
385	366.0	censored
387	28.3	lapse
391	45.1	lapse
393	338.9	censored
398	75.8	lapse

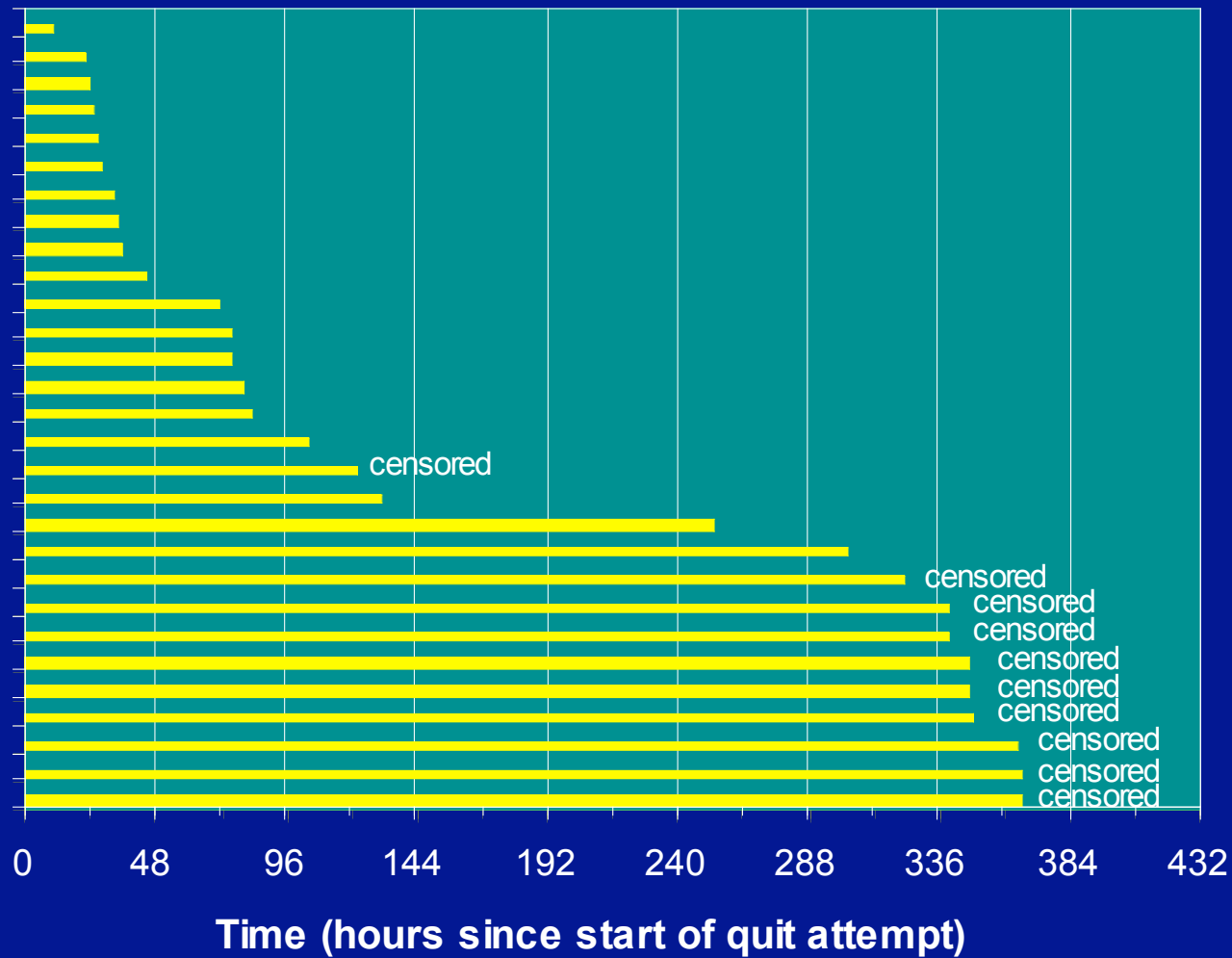
## Plot of raw data, distinguishing first lapses from censored observations



## Data sorted by Hours

<u>ID</u>	<u>Hours</u>	<u>Status</u>
373	10.5	lapse
354	21.6	lapse
331	23.8	lapse
352	24.3	lapse
309	26.1	lapse
387	28.3	lapse
337	32.4	lapse
384	33.3	lapse
359	34.8	lapse
391	45.1	lapse
376	70.9	lapse
365	75.5	lapse
398	75.8	lapse
307	80.0	lapse
364	82.4	lapse
326	103.3	lapse
318	122.4	censored
338	130.6	lapse
333	252.7	lapse
328	302.4	lapse
380	322.4	censored
393	338.9	censored
341	340.0	censored
325	346.8	censored
366	346.9	censored
370	347.6	censored
327	365.3	censored
385	366.0	censored
329	366.1	censored

## Sorted raw data, distinguishing first lapses from censored observations



Can begin to estimate  
survival function

But, what do we do  
with censored  
observations?

<u>ID</u>	<u>Hours</u>	<u>Status</u>	<u>Cum Pct</u>	<u>% Surviving</u>
373	10.5	lapse	3.4%	96.6%
354	21.6	lapse	6.9%	93.1%
331	23.8	lapse	10.3%	89.7%
352	24.3	lapse	13.8%	86.2%
309	26.1	lapse	17.2%	82.8%
387	28.3	lapse	20.7%	79.3%
337	32.4	lapse	24.1%	75.9%
384	33.3	lapse	27.6%	72.4%
359	34.8	lapse	31.0%	69.0%
391	45.1	lapse	34.5%	65.5%
376	70.9	lapse	37.9%	62.1%
365	75.5	lapse	41.4%	58.6%
398	75.8	lapse	44.8%	55.2%
307	80.0	lapse	48.3%	51.7%
364	82.4	lapse	51.7%	48.3%
326	103.3	lapse	55.2%	44.8%
318	122.4	censored	58.6%	
338	130.6	lapse	62.1%	
333	252.7	lapse	65.5%	
328	302.4	lapse	69.0%	
380	322.4	censored	72.4%	
393	338.9	censored	75.9%	
341	340.0	censored	79.3%	
325	346.8	censored	82.8%	
366	346.9	censored	86.2%	
370	347.6	censored	89.7%	
327	365.3	censored	93.1%	
385	366.0	censored	96.6%	
329	366.1	censored	100.0%	

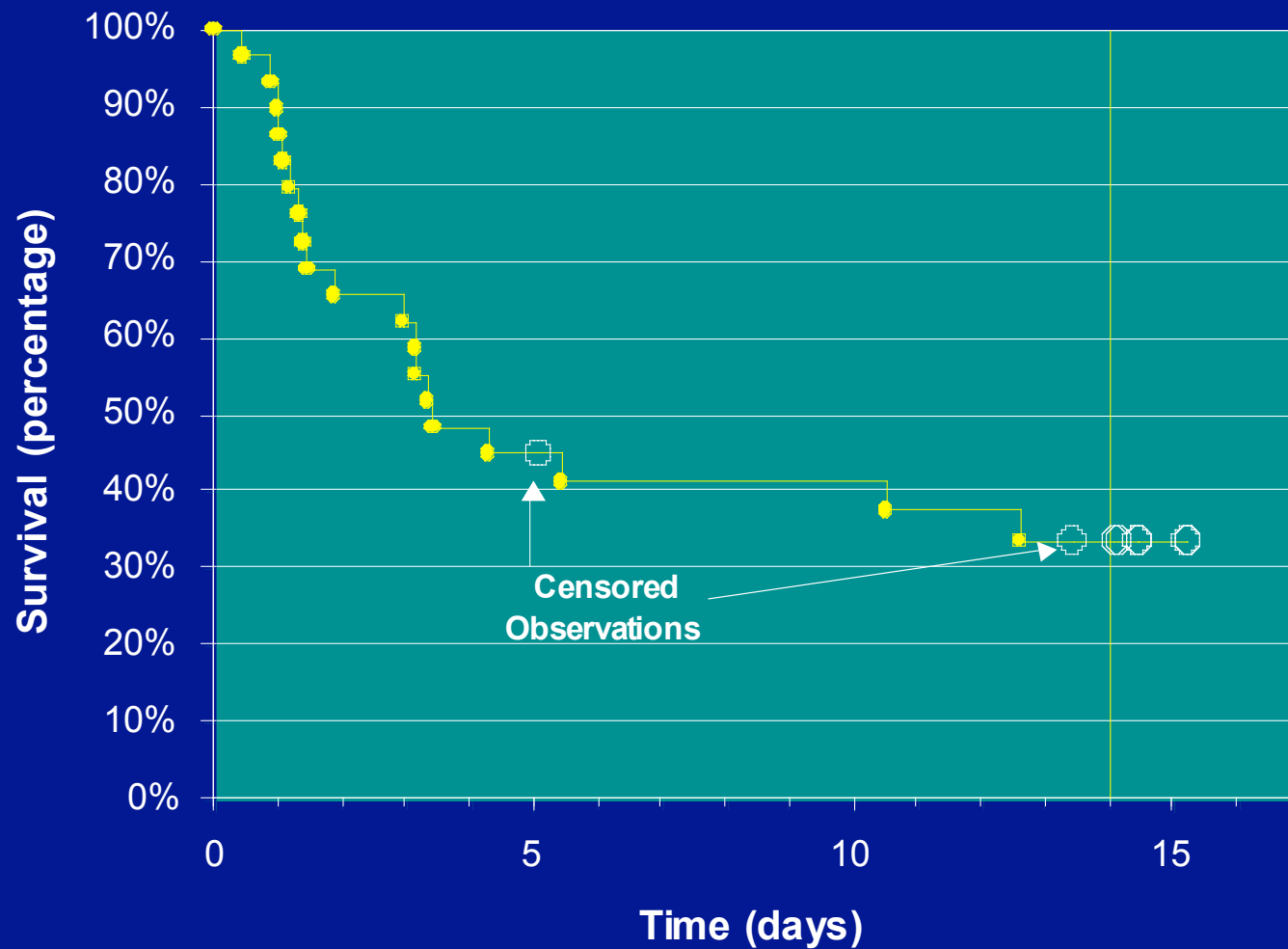
# Kaplan-Meier estimate of survival function/ distribution

<u>ID</u>	<u>Hours</u>	<u>Status</u>	<u>Cum Pct</u>	<u>% Surviving</u>	
373	10.5	lapse	3.4%	96.6%	= 1.00 (1 - 1/29)
354	21.6	lapse	6.9%	93.1%	= .966 (1 - 1/28)
331	23.8	lapse	10.3%	89.7%	= .931 (1 - 1/27)
352	24.3	lapse	13.8%	86.2%	= .897 (1 - 1/26)
309	26.1	lapse	17.2%	82.8%	= .862 (1 - 1/25)
387	28.3	lapse	20.7%	79.3%	= .828 (1 - 1/24)
337	32.4	lapse	24.1%	75.9%	= .793 (1 - 1/23)
384	33.3	lapse	27.6%	72.4%	= .759 (1 - 1/22)
359	34.8	lapse	31.0%	69.0%	= .724 (1 - 1/21)
391	45.1	lapse	34.5%	65.5%	= .690 (1 - 1/20)
376	70.9	lapse	37.9%	62.1%	= .655 (1 - 1/19)
365	75.5	lapse	41.4%	58.6%	= .621 (1 - 1/18)
398	75.8	lapse	44.8%	55.2%	= .586 (1 - 1/17)
307	80.0	lapse	48.3%	51.7%	= .552 (1 - 1/16)
364	82.4	lapse	51.7%	48.3%	= .517 (1 - 1/15)
326	103.3	lapse	55.2%	44.8%	= .483 (1 - 1/14)
<b>318</b>	<b>122.4</b>	<b>censored</b>	<b>58.6%</b>	<b>(44.8%)</b>	
<b>338</b>	<b>130.6</b>	<b>lapse</b>	<b>62.1%</b>	<b>41.1%</b>	<b>= .448 (1 - 1/12)</b>

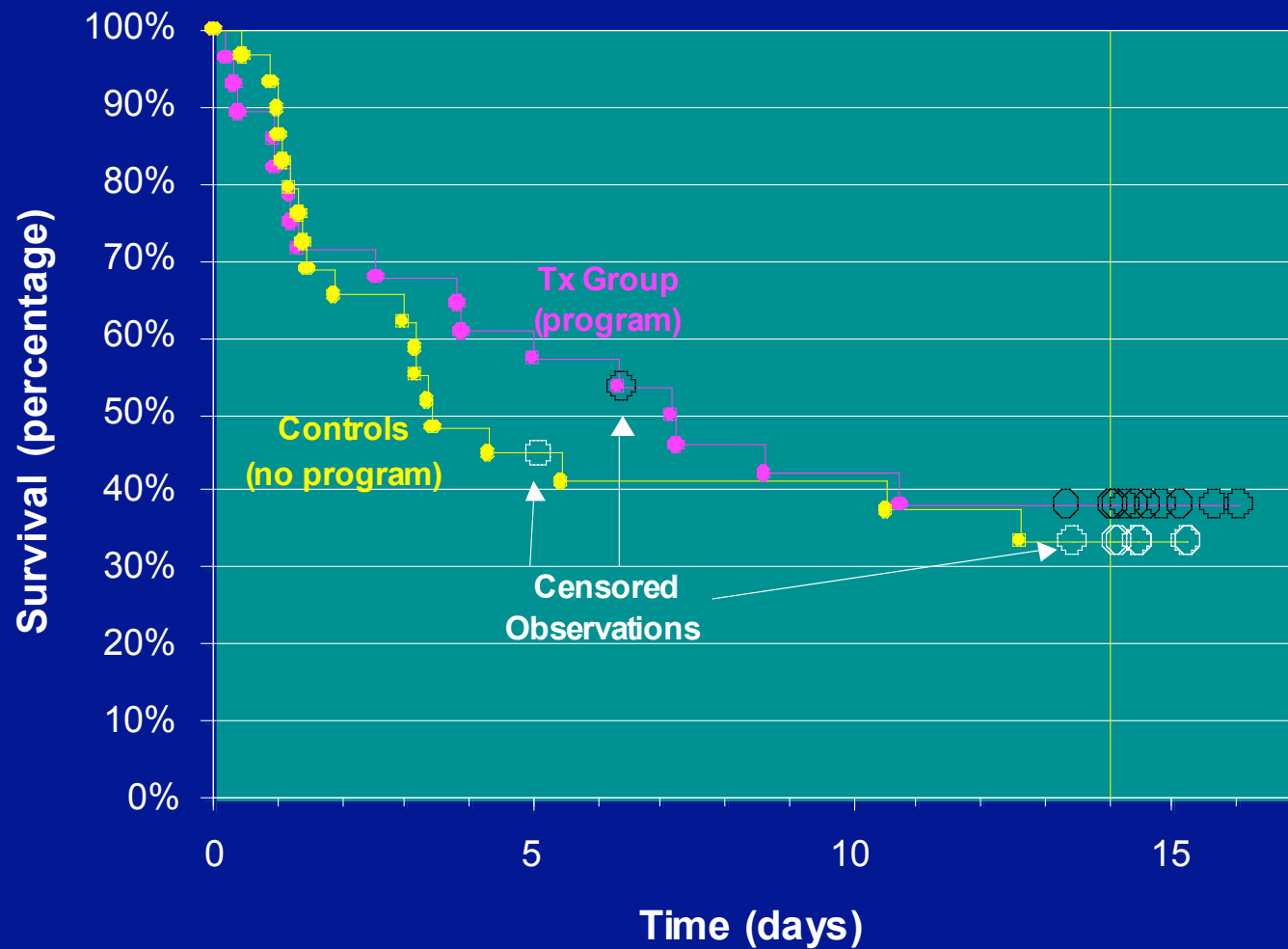
# Kaplan-Meier estimate of survival function/ distribution

<u>ID</u>	<u>Hours</u>	<u>Status</u>	<u>Cum Pct</u>	<u>% Surviving</u>	
373	10.5	lapse	3.4%	96.6%	= 1.00 (1 - 1/29)
354	21.6	lapse	6.9%	93.1%	= .966 (1 - 1/28)
331	23.8	lapse	10.3%	89.7%	= .931 (1 - 1/27)
352	24.3	lapse	13.8%	86.2%	= .897 (1 - 1/26)
309	26.1	lapse	17.2%	82.8%	= .862 (1 - 1/25)
387	28.3	lapse	20.7%	79.3%	= .828 (1 - 1/24)
337	32.4	lapse	24.1%	75.9%	= .793 (1 - 1/23)
384	33.3	lapse	27.6%	72.4%	= .759 (1 - 1/22)
359	34.8	lapse	31.0%	69.0%	= .724 (1 - 1/21)
391	45.1	lapse	34.5%	65.5%	= .690 (1 - 1/20)
376	70.9	lapse	37.9%	62.1%	= .655 (1 - 1/19)
365	75.5	lapse	41.4%	58.6%	= .621 (1 - 1/18)
398	75.8	lapse	44.8%	55.2%	= .586 (1 - 1/17)
307	80.0	lapse	48.3%	51.7%	= .552 (1 - 1/16)
364	82.4	lapse	51.7%	48.3%	= .517 (1 - 1/15)
326	103.3	lapse	55.2%	44.8%	= .483 (1 - 1/14)
318	122.4	censored	58.6%	(44.8%)	
338	130.6	lapse	62.1%	41.1%	= .448 (1 - 1/12)
333	252.7	lapse	65.5%	37.4%	= .411 (1 - 1/11)
328	302.4	lapse	69.0%	33.6%	= .387 (1 - 1/10)
380	322.4	censored	72.4%	(33.6%)	
393	338.9	censored	75.9%	(33.6%)	
341	340.0	censored	79.3%	(33.6%)	
325	346.8	censored	82.8%	(33.6%)	
366	346.9	censored	86.2%	(33.6%)	
370	347.6	censored	89.7%	(33.6%)	
327	365.3	censored	93.1%	(33.6%)	
385	366.0	censored	96.6%	(33.6%)	
329	366.1	censored	100.0%	(33.6%)	

## "Success" during First 2 Weeks of Smoking Cessation Attempt, Control Group (K-M curve)



## "Success" during First 2 Weeks of Smoking Cessation Attempt, Kaplan-Meier Curves



# Censoring and Kaplan-Meier

- Kaplan-Meier method allows us to estimate the distribution of survival times
  - It handles censored observations by utilizing the information that no event occurred prior to the time of censoring, while recognizing that we do not know if or when the event occurred after this time.
  - Assumes the censoring is “non-informative,” i.e., that the occurrence/timing of censored observations is unrelated to their risk of an event

- Examples where censoring likely to be non-informative:
  - Participant moved because company moved
  - Participant killed by lightning
  - Administrative censoring (trial ended)
  - Study equipment failed
- Examples where censoring may be *informative*:
  - Participant withdrew or dropped out of sight, perhaps because she was about to begin smoking

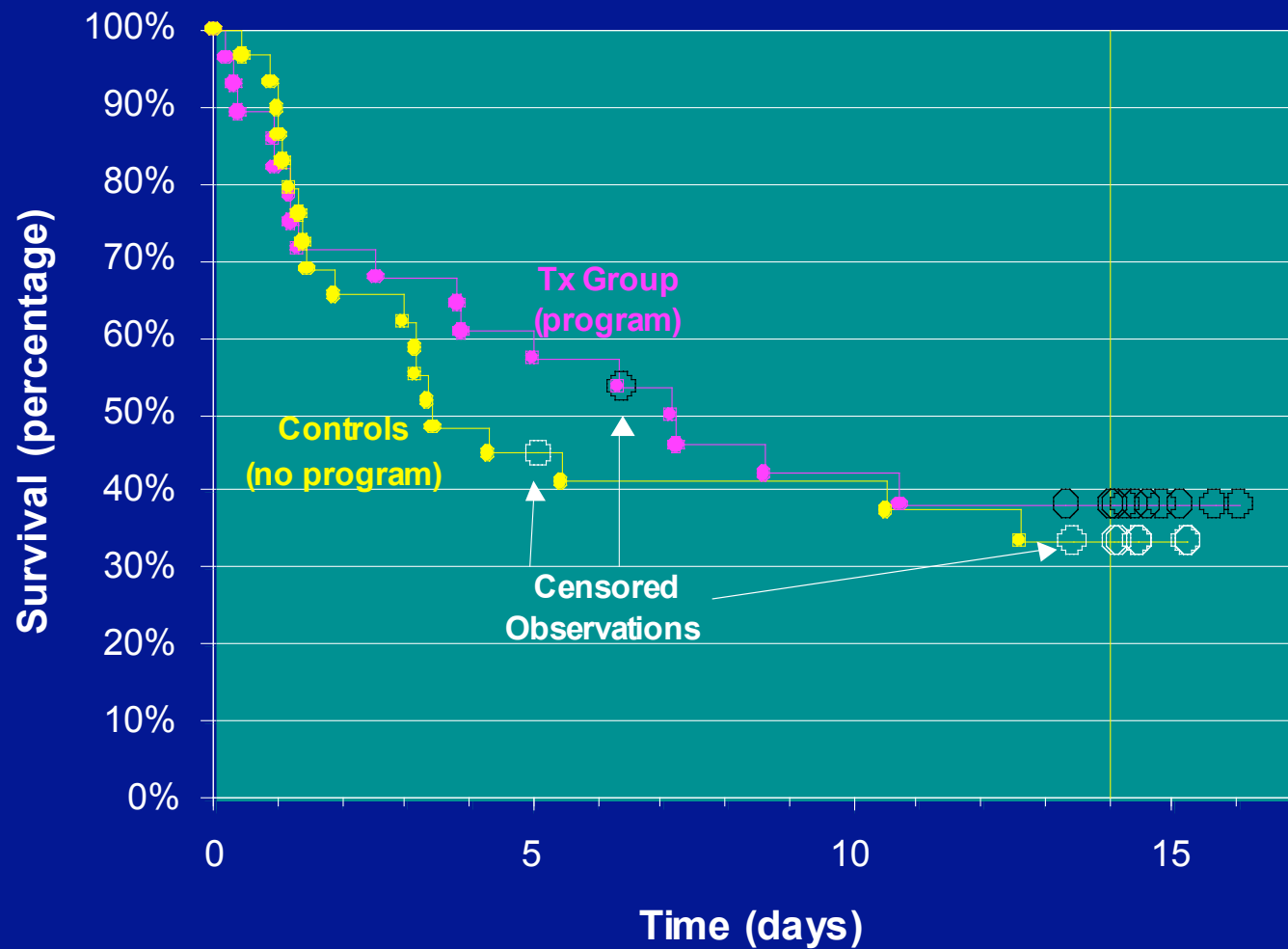
# Hazard ratio

- We often test whether one group has a higher mean than the other (assuming that the shape of the distribution is similar in the two groups; i.e., normal with the same s.d.).
- In logistic regression, we test whether the odds of a binary outcome differs between two groups (odds ratio differs from 1.0)

# Hazard ratio

- In survival analysis we most often test for differences in survival by assuming that the hazard function  $\lambda(t)$  in one group is proportional to that in the other and testing whether  $\lambda_{Tx}(t) / \lambda_C(t)$  differs from 1.0.
- $\lambda_{Tx}(t) / \lambda_C(t)$  is called the “hazard ratio” and is very closely related to the concept of relative risk.
- Note: when proportional hazards assumption is true, we do not expect survival curves to cross (except by chance)

## "Success" during First 2 Weeks of Smoking Cessation Attempt, Kaplan-Meier Curves



# Testing the difference between two survival distributions

- Non-parametric tests of whether one distribution is “reliably” above/below other
  - Log-rank test (Mantel)
    - most powerful when two distributions differ by a fixed proportion ( $\lambda_{Tx}(t) / \lambda_C(t) = \text{constant}$ )
    - gives greater weight to later survival times
  - Wilcoxon test (Gehan)
    - gives greater weight to earlier events
  - Each yields a chi-square statistic w/  $df=1$

# Did social support tx delay resumption of smoking?

Test	Chi-Square	DF	Pr >
			Chi-Square
Log-Rank	0.2024	1	0.6528
Wilcoxon	0.1494	1	0.6991

Answer: group difference in survival curves  
does not approach statistical significance  
(perhaps study underpowered)

# Testing the difference between two survival distributions

- Non-parametric tests of whether the  $p$ -th percentile (e.g. median) differs in the two distributions
  - each survival estimate for the survival distribution has an associated standard error/variance
  - these can be used to create a chi-square or z-test of equality
  - issue: choice of percentile(s)

# Models that adjust for covariates

- Stratified analyses
  - estimate separate survival distributions for each stratum of a 3rd variable and test for treatment/control differences within each stratum (issue: how to pool across strata)
- Hazard regression analysis (Cox proportional hazards model)
  - regression-like model that partials out the effects of one or more 3rd variables
  - assumes each variable has multiplicative effect on survival

# Cox proportional hazards regression analysis

<u>Variable</u>	<u>DF</u>	Parameter <u>Estimate</u>	Standard <u>Error</u>
Group	1	-0.15040	0.33478

<u>Hazard Ratio</u>	<u>95% Hazard Ratio Confidence Limits</u>	
0.860	0.446	1.658

$\lambda_{Tx}(t) / \lambda_C(t) = .86$  (ns)  
(likelihood ratio test equivalent to log-rank test)

<u>Variable</u>	<u>DF</u>	<u>Parameter Estimate</u>	<u>Standard Error</u>
Group	1	-0.15301	0.33535
Sex	1	-0.05020	0.36271

<u>Hazard Ratio</u>	<u>95% Hazard Ratio Confidence Limits</u>	
0.858	0.445	1.656
0.951	0.467	1.936

$$\lambda_{Tx}(t) / \lambda_C(t) = .86 \text{ (ns)}$$

# Time-varying covariates

- Can be used to handle covariates that change over time
  - use of nicotine patch/gum on some days
  - whether participant took medication
  - changes in participant adherence to protocol
  - weekdays vs weekends

# Concepts covered

1. Survival curve/function
2. Hazard rate
3. Censored data
4. Kaplan-Meier estimate of survival curve
5. Hazard ratio
6. Proportional hazards assumption
7. Comparing survival curves
8. Proportional hazards regression analysis
9. Time-varying covariates